ESTIMATION OF THE P(CS) WITH A COMPUTER PROGRAM

by

Edward J. Dudewicz*

ESTIMATION OF THE P(CS) WITH A COMPUTER PROGRAM

by

Edward J. Dudewicz*

Technical Report No. 195a
Department of Statistics
The Ohio State University
Columbus, Ohio 43210
December 1979 (Revised July 1980)

* This author's research was supported by Office of Naval Research Contract No. N00014-78-C-0543.

DISCLAIMER NOTICE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: TR-195a
2. GOVT ACCESSION NO.: AD-A088058
3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): ESTIMATION OF THE P(CS) WITH A COMPUTER PROGRAM
5. TYPE OF REPORT & PERIOD COVERED: Technical Report
6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): Edward J. Dudewicz
8. CONTRACT OR GRANT NUMBER(s): N00014-78-C-0543
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Statistics, The Ohio State University, Columbus, Ohio 43210
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NR 042-403
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research, Department of the Navy, Arlington, Virginia 22217
12. REPORT DATE: July [redacted]80
13. NUMBER OF PAGES: ii + 12
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office):
15. SECURITY CLASS. (of this report): Unclassified
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:
16. DISTRIBUTION STATEMENT (of this Report): Approval for public release; distribution unlimited.
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report):
18. SUPPLEMENTARY NOTES:
19. KEY WORDS (Continue on reverse side if necessary and identify by block number): Key Words and Phrases: Ranking and selection; estimation of P(CS); probability of correct selection; inequalities; Monte Carlo; sequential; adaptive; heteroscedastic.
20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

In the area of ranking and selection, recent papers have presented procedures which attempt to allow experimenters to take at least partial advantage of more-favorable configurations of the population parameters (compared to least-favorable configuration). The current paper presents a computer program which allows experimenters to dispense with all of inequalities, an indiffernece-zone, and known equal variabilities in their assessment of the probability of correct selection. It has additional uses in implementation of recently-proposed sequential adaptive selection procedures.

DD FORM 1 JAN 73 1473

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ESTIMATION OF THE P(CS) WITH A COMPUTER PROGRAM*

by

Edward J. Dudewicz

ABSTRACT

In the area of ranking and selection, recent papers have presented procedures which attempt to allow experimenters to take at least partial advantage of more-favorable configurations of the population parameters (compared to the least-favorable configuration). The current paper presents a computer program which allows experimenters to dispense with all of inequalities, an indifference-zone, and known equal variabilities in their assessment of the probability of correct selection. It has additional uses in implementation of recently-proposed sequential adaptive selection procedures.

* This research was planned and initiated while Edward J. Dudewicz was a Visiting Scholar, Department of Mathematics and Statistics, University of Nebraska, Lincoln. Edward J. Dudewicz is Professor, Department of Statistics, The Ohio State University, Columbus, Ohio 43210, U.S.A.

This research was supported by Office of Naval Research Contract No. N00014-78-C-0543.

ESTIMATION OF THE P(CS) WITH A COMPUTER PROGRAM

by

Edward J. Dudewicz

1. INTRODUCTION

When an experimenter uses the indifference-zone ranking-and-selection procedures pioneered by Bechhofer (1954) and recently explicated by Gibbons, Olkin, and Sobel (1977), he sets his sample size(s) to guarantee a P(CS) (probability of correct selection) at least P^* whenever a minimal separation (of the best population from the other populations) of at least δ^* exists; a separation of δ^* is then least-favorable. The actual separation is likely to be more favorable than least-favorable (and the P(CS) thus $> P^*$), and experimenters would like to take advantage of this favorableness after the experiment has been run. Previous approaches to this problem by Olkin, Sobel, and Tong (1976) and Anderson, Bishop, and Dudewicz (1977) all involved inequalities. In this paper we give a computer program which allows assessment of the P(CS) without use of inequalities, is simple and easily implemented on even small computers, and does not use an indifference-zone. It is also directly useful when (as is often the case) the samples are drawn before the statistician is consulted, and he must evaluate their adequacy. It has additional uses in implementation of recently proposed sequential adaptive selection procedures of Tong (1978).

While our approach is oriented towards the practitioner and computer software provider in this paper, it should be noted that the area is also one of recent intensive theoretical investigation by Olkin, Sobel, and Tong (1979), Bofinger (1980), and Faltin (1980 a, b).

Note that Olkin, Sobel, and Tong (1979) consider only the case of σ^2 known. Work of Bofinger (1980) contradicts their result on asymptotic normality of their $P(CS)$ estimator. Faltin (1980a) shows this estimator is biased on the high side when $k = 2$, while Faltin (1980b) gives an alternate quantile unbiased estimator (also for the case $k = 2$). By contrast, our estimator allows for unknown and heteroscedastic variances and does not require either bounds or tables: the experimenter makes one inexpensive run of a simple computer program (presented in full detail with numerical examples below).

2. THE P(CS) FUNCTION

If the experimenter has obtained a random sample of size n_i from population π_i which yields normally distributed observations with unknown mean μ_i and variance σ_i^2, $1 \le i \le k$, and wishes to evaluate the adequacy of these sample sizes when the means procedure ("select the population yielding the largest sample mean as having the largest population mean") is used (and all $n_1 + \ldots + n_k$ observations are independent), he will need to assess the function

$$P(CS) = \int_{-\infty}^{\infty} \left\{ \prod_{i=1}^{k-1} \Phi\left(\frac{\sigma_{(k)}}{\sigma_{(i)}} \sqrt{\frac{n_{(i)}}{n_{(k)}}} \; z + \frac{\mu_{[k]} - \mu_{[i]}}{\sigma_{(i)}/\sqrt{n_{(i)}}} \right) \right\} \phi(z)dz$$

where: $\mu_{[1]} \le \cdots \le \mu_{[k]}$ denote $\mu_1, \ldots, \mu_k$ in numerical order; $\sigma_{(i)}$, $n_{(i)}$ are the standard deviation and sample size of the population with true mean $\mu_{[i]}$; Φ, ϕ are the standard normal distribution function and density function.

Note that, with error ϵ ($0 < \epsilon < .005$ since $\Phi(2.81) = .9975$),

$$P(CS) = \int_{-2.81}^{2.81} \left\{ \prod_{i=1}^{k-1} \Phi\left(\frac{\sigma_{(k)}}{\sigma_{(i)}} \sqrt{\frac{n_{(i)}}{n_{(k)}}} \; z + \frac{\mu_{[k]} - \mu_{[i]}}{\sigma_{(i)}/\sqrt{n_{(i)}}} \right) \right\} \phi(z)dz.$$

Once the $\mu_{[i]}$, $\sigma_{(i)}$ are estimated from the samples, this can be estimated simply by Monte Carlo methods on even small computers, with ease compared to quadrature methods, and with accuracy compared to methods involving inequalities. A computer program is given in Section 3, with an example in Section 4.

3. COMPUTER PROGRAM

The following computer program reads K (the number of populations) and XM(I), S(I), N(I) (the sample mean, sample standard deviation, and sample size of population I), $1 \leq I \leq K$, from data cards and then uses 40,000 Monte Carlo trials to estimate P(CS) with a standard error not exceeding .005/2, so that (with 95% confidence, also taking account of the truncation error ϵ) we estimate the P(CS) integral within .01. A numerical example is given in section 4.

```
C
C     THE FOLLOWING COMPUTER PROGRAM:
C        1.  READS K (THE NUMBER OF POPULATIONS) FROM A DATA CARD;
C        2.  READS XM(I),S(I),N(I) (THE SAMPLE MEAN, STANDARD DEVIATION
C              AND SIZE OF POPULATION I),  1<=I<=K, FROM K DATA CARDS;
C        3.  USES 40,000 MONTE CARLO TRIALS TO ESTIMATE THE P(CS) WITH
C              A STANDARD ERROR NOT EXCEEDING .005/2, SO THAT (WITH 95
C              PERCENT CONFIDENCE, ALSO TAKING ACCOUNT OF THE
C              TRUNCATION ERROR) WE ESTIMATE THE P(CS) INTEGRAL WITHIN
C              .01.
C
      INTEGER R,SEED
      DIMENSION XM(50),S(50),N(50),R(50)
C
C         PERSONS WITH K>50 SHOULD INCREASE ALL FOUR DIMENSIONS ABOVE
C         TO THAT NUMBER BEFORE RUNNING THIS PROGRAM.  THEY SHOULD
C         ALSO INCREASE THE DIMENSIONS OF YM(.) AND S(.) IN SUB-
C         ROUTINE VSORT.
C
      READ(5,20) K
   20 FORMAT(I10)
      DO 30 I = 1,K
      READ(5,21) XM(I),S(I),N(I)
   21 FORMAT(F10.0,F10.0,I10)
   30 CONTINUE
C
C         WE NOW FIND INTEGERS R(1),...,R(K) SUCH THAT
C         XM(R(1))<=XM(R(2))<=...<=XM(R(K)) BY A VIRTUAL SHELL SORT.
C
      CALL VSORT(XM,R,K)
C
      SEED=967021
      I=0
      JCNT=0
C
C         WE NOW ENTER THE MONTE CARLO LOOP.
C
   31 CONTINUE
      IF(I.GE.40000)GO TO 34
      CALL URN20(SEED,RANX,1)
      CALL URN20(SEED,RANY,1)
      RANX=(RANX-0.5)*2.0*2.81
      RANY=RANY*(1.0/SQRT(2.0*3.14159))
C
      PROD=1.0
      KK = K-1
      DO 32 L=1,KK
      TERM=XM(R(K))-XM(R(L))
      TERM=TERM/S(R(L))
      AN = FLOAT(N(R(L)))
      TERM = TERM*SQRT(AN)
      FACT=S(R(K))/S(R(L))
      CN = FLOAT(N(R(L)))
      DN = FLOAT(N(R(K)))
      BN = CN/DN
      FACT = FACT*SQRT(BN)
      FACT=FACT*RANX+TERM
      TERM=DCDFN(FACT)
```

```
      PROD = PROD*TERM
   32 CONTINUE
      FN=PROD/SQRT(2.0*3.14159)
      FN=FN*EXP(-0.5*RANX*RANX)
      IF(FN.LE.RANY)GO TO 33
      JCNT=JCNT+1
   33 CONTINUE
      I=I+1
      GO TO 31
C
C          WE HAVE DONE 40000 MONTE CARLO TRIALS, JCNT OF WHICH YIELDED
C          A VALUE OF THE INTEGRAND ABOVE RANY.  WE NOW CALCULATE AND
C          REPORT THE P(CS) ESTIMATE AND ITS STANDARD DEVIATION.
C
   34 CONTINUE
      PCSEST=(JCNT/40000.)*2.81*2.0*(1.0/SQRT(2.0*3.14159))
      STDER=(JCNT/40000.)*(1.0-(JCNT/40000.))
      STDER=SQRT(STDER)/200.
      STDER=STDER*2.81*2.*(1.0/SQRT(2.0*3.14159))
C
      WRITE(6,41)
   41 FORMAT(1H1)
      WRITE(6,42)
   42 FORMAT(1H0,6X,54HTHE PROBABILITY OF CORRECT SELECTION IN A PROBLEM
     1 WITH)
      WRITE(6,43) K
   43 FORMAT(1H0,6X,5H  K= ,I2,12H POPULATIONS)
      DO 35  I=1,K
      WRITE(6,44)  I,XM(I),S(I),N(I)
   44 FORMAT(1H0,6X,13H  POPULATION ,I2,10H HAS MEAN ,F10.4,
     112H, STD. DEV.  ,F10.4,19H, AND SAMPLE SIZE  ,I4)
   35 CONTINUE
      WRITE(6,45) PCSEST
   45 FORMAT(1H0,6X,11HIS P(CS) = ,F6.4,34H BASED ON 40000 MONTE CARLO T
     1RIALS)
      WRITE(6,46) STDER
   46 FORMAT(1H0,6X,25HAND HAS STANDARD ERROR = ,F6.4)
      WRITE(6,47)
   47 FORMAT(1H1)
      END
```

```
      SUBROUTINE VSORT(XM,R,K)
      INTEGER R,S,SAV2
      DIMENSION XM(1),R(1),YM(50),S(50)
      DO 10 I=1,K
      YM(I)=XM(I)
      S(I)=I
   10 CONTINUE
C
      M=K
    1 M=M/2
      IF(M.EQ.0)GO TO 5
      J=1
    2 I=J
    4 IF(YM(I).LE.YM(I+M)) GO TO 3
      SAV1=YM(I)
      SAV2=S(I)
      YM(I)=YM(I+M)
      S(I)=S(I+M)
      YM(I+M)=SAV1
      S(I+M)=SAV2
      I=I-M
      IF(I.GE.1) GO TO 4
    3 J=J+1
      IF(J.GT.K-M) GO TO 1
      GO TO 2
    5 CONTINUE
      DO 11 I=1,K
      R(I) = S(I)
   11 CONTINUE
      RETURN
      END
```

```
C
C     FUNCTION DCDFN(X)
C
C*****************************************************************************
C
C A SUBROUTINE TO CALCULATE PROBABILITIES FOR STANDARD NORMAL RANDOM
C    VARIABLES.
C
C PURPOSE
C    GENERATES THE PROBABILITY THAT Z IS LESS THAN OR EQUAL TO X,
C    WHERE Z IS A STANDARD NORMAL RANDOM VARIABLE
C
C USAGE
C    "LET P=DCDFN(X)"
C
C DESCRIPTION OF PARAMETERS
C    X...IS THE CUTOFF POINT FOR WHICH THE PROBABILITY THAT Z IS LESS
C         THAN OR EQUAL TO X IS DESIRED
C
C REMARKS
C    THIS IS A DOUBLE PRECISION FUNCTION AND CAN ACCOMMODATE A DOUBLE
C    PRECISION ARGUMENT
C
C SUBROUTINES AND FUNCTION SUBPROGRAMS REQUIRED
C    NONE
C
C METHOD
C    FOR THE ALGORITHM USED AND ITS ACCURACY, SEE REFERENCE 2.  FOR
C    SPEED AND COMPARISONS WITH OTHER ALGORITHMS (WITH REGARD TO SPEED
C    AND ACCURACY), SEE REFERENCE 1.
C    1. DUDEWICZ,E.J. AND RALLEY,T.: "TES[RAND-RANDOM NUMBER GENERATION
C       AND TESTING PACKAGE (INCLUDING IRCCRAND)," TECHNICAL REPORT,
C       DEPARTMENT OF STATISTICS, THE OHIO STATE UNIVERSITY, COLUMBUS,
C       OHIO. 1979.
C    2. HILTON,R.C. AND HOTCHKISS,R.: "COMPUTER EVALUATION OF THE
C       NORMAL AND INVERSE NORMAL DISTRIBUTION FUNCTIONS,"
C       TECHNOMETRICS, VOL.11(1969), PP. 817-822.
C
C*****************************************************************************
C
      DOUBLE PRECISION FUNCTION DCDFN(X)
      DIMENSION A(6),D(6),C(6,19)
      REAL*8 A,C,D,X,Y,Z,SGNY
      DATA A/.625D0,1.25D0,2.0D0,2.45D0,3.5D0,4.62D0/
      DATA D/.3D0,.925D0,1.625D0,2.225D0,2.95D0,4.15D0/
      DATA C/6.7932493291D-04,-1.2709753598D-03,6.7964525797D-04,
     *-8.8509314297D-05,1.4791554152D-07,5.4379072878D-07,
     *5.1925660850D-03,-1.4822549744D-03,-3.4539893722D-04,
     *5.6326070991D-04,-6.1965913125D-05,-8.6656125997D-07,
     *-6.0189652379D-03,8.7718956094D-03,-3.0853491043D-03,
     *-4.5015531923D-04,3.2265719341D-04,6.6377769441D-07,
     *-2.2756113780D-02,-2.0233415601D-03,7.2141082997D-03,
     *-1.3672309725D-03,-4.4397018986D-04,-6.4904533798D-07,
     *3.3558772127D-02,-3.4631674408D-02,-1.0677959140D-03,
     *5.5422166916D-03,6.2973759838D-04,6.2471724778D-07,
     *7.2703718850D-02,4.7642392207D-02,-2.4859760554D-02,
     *-1.0221129383D-02,-6.6385580832D-04,-4.2105904094D-07,
     *-1.4970895354D-01,5.6850273056D-02,5.7420315819D-02,
```

Best Available Copy

```
     *1.1850195831D-02,5.1265578121D-04,2.0742133303D-07,
     *-1.5466912979D-01,-2.2180615138D-01,-6.5383705170D-02,
     *-9.8864030978D-03,-2.7637162532D-04,-7.6850942218D-08,
     *5.1563045460D-01,2.3979043973D-01,4.0236129479D-02,
     *6.9950835701D-03,9.3751413427D-05,1.8722029323D-08,
     *1.6481387974D-01,4.0458335515D-01,4.8922185659D-01,
     *4.9917416703D-01,4.9998489849D-01,4.9999999779D-01/
      Y=X*0.70710678119D0
      SGNY=1.D0
      IF(Y)2,1,3
    1 DCDFN=.5D0
      RETURN
    2 SGNY=-1.D0
      Y=-Y
    3 DO 4 I=1,6
      IF(Y-A(I))5,5,4
    4 CONTINUE
      Z=.5D0
      GO TO 7
    5 Y=Y-B(I)
      Z=C(I,1)
      DO 6 J=2,10
    6 Z=Z*Y+C(I,J)
    7 DCDFN=.5D0+SGNY*Z
      RETURN
      END
```

```
URN20    CSECT
         USING URN20,15
         STM   14,12,12(13)
         ST    13,SAVE+4
         LR    11,13              COPY ADDR OF CALLING SAVE INTO REG 11
         LA    13,SAVE            ADDR OF SAVE IN REG13
         ST    13,8(11)           STORE ADDR OF SAVE IN CALLING PROC SAVE
*
         LM    2,4,0(1)           REG(2)=ADDR(IX) REG(3)=ADDR(X)
*                                 REG(4)=ADDR(NBATCH)
         L     11,0(4)            VALUE OF NBATCH IN REG11
         L     10,=F'0'           INIT INDEX
         L     9,=X'40000000'     EXPONENT FOR FLOATING POINT CONVERSION
LOOP     L     7,0(2)             VALUE OF IX IN REG7
         M     6,MLT              REG6=QUOT, REG7=REM  MOD 2**32
         SLDL  6,1                PUTS QUOT MOD 2**31 IN REG6
         SRL   7,1                PUTS REM  MOD 2**31 IN REG7
         AR    6,7                REG6 = REM MOD 2**31-1=
*                                   REM MOD 2**31 + QUOT MOD 2**31
         C     6,=F'0'            CHECK FOR OVERFLOW
         BNL   UP1                BRANCH TO UP1 IF SUM IS NOT NEG
         SLL   6,1                DIVIDE REG6 BY 2**31
         SRL   6,1                REM MOD 2**31 IN REG6
         A     6,=F'1'            ADD QUOT FROM DIVISION TO REM TO
*                                   OBTAIN REM MOD 2**31-1 IN REG6
UP1      ST    6,0(2)             STORE VALUE OF IX FOR NEXT RND NOS
         LPR   6,6                ABSOL VALUE OF IX IN REG6
         L     5,=X'4E000000'     EXPONENT IN REG5 (16**14)
         STM   5,6,DOUBLE         STORE FLT PT NOS IN REGS5 & 6
         LD    2,DOUBLE           LOAD FLT PT REG2 WITH FLT PT NOS
         AD    2,=D'0.0'          NORMALIZE BY ADDING 0.
         DD    2,MODUL              DIVIDE BY 2**31-1
         STD   2,0(10,3)          STORE NORMALIZED FLT PT NOS IN X(INDEX)
         LA    10,4(10)           INCREMENT INDEX
         BCT   11,LOOP            BOTTOM OF LOOP
         L     13,SAVE+4          RESTORE REGISTERS
         LM    14,12,12(13)
         BCR   15,14              RETURN TO CALLING PROGRAM
MLT      DC    F'2027812808'      MULTIPLIER FOR URN20
DOUBLE   DS    D
MODUL    DC    X'487FFFFFFF000000'
SAVE     DS    18F
         END
               =D'0.0'
               =F'0'
               =X'40000000'
               =F'1'
               =X'4E000000'
```

4. EXAMPLE AND COMPARISON

As an example, when we have k = 8 populations with means all 0.0 except for one which is 2.34, with standard deviations all 9.0, and sample sizes all 81, we know (from the tables of Bechhofer (1954)) that the true P(CS) = .80. Our computer program yields the output

```
THE PROBABILITY OF CORRECT SELECTION IN A PROBLEM WITH
  K=  8 POPULATIONS
  POPULATION  1 HAS MEAN     2.3400, STD. DEV.     9.0000, AND SAMPLE SIZE    81
  POPULATION  2 HAS MEAN     0.0   , STD. DEV.     9.0000, AND SAMPLE SIZE    81
  POPULATION  3 HAS MEAN     0.0   , STD. DEV.     9.0000, AND SAMPLE SIZE    81
  POPULATION  4 HAS MEAN     0.0   , STD. DEV.     9.0000, AND SAMPLE SIZE    81
  POPULATION  5 HAS MEAN     0.0   , STD. DEV.     9.0000, AND SAMPLE SIZE    81
  POPULATION  6 HAS MEAN     0.0   , STD. DEV.     9.0000, AND SAMPLE SIZE    81
  POPULATION  7 HAS MEAN     0.0   , STD. DEV.     9.0000, AND SAMPLE SIZE    81
  POPULATION  8 HAS MEAN     0.0   , STD. DEV.     9.0000, AND SAMPLE SIZE    81
IS P(CS) = 0.8030 BASED ON 40000 MONTE CARLO TRIALS
AND HAS STANDARD ERROR = 0.0034
```

which agrees with the theoretical value, thus furnishing a check on the computer program.

REFERENCES

Anderson, P. O., Bishop, T. A., and Dudewicz, E. J. (1977). Indifference-zone ranking and selection: confidence intervals for true achieved P(CD). Communications in Statistics - Theory and Methods, Vol. A6, pp. 1121-1132.

Bechhofer, R. E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. Annals of Mathematical Statistics, Vol. 25, pp. 16-39.

Bofinger, E. (1980). On the non-existence of consistent estimators for P{C.S.}. Unpublished manuscript.

Faltin, F. W. (1980a). Performance of the Sobel-Tong estimator of the probability of correct selection achieved by Bechhofer's single-stage procedure for the normal means problem. Abstract, IMS Bulletin, Vol. 9, p. 180.

Faltin, F. W. (1980b). A quantile unbiased estimator of the probability of correct selection achieved by Bechhofer's single-stage procedure for the two population normal means problem. Abstract, IMS Bulletin, Vol. 9, pp. 180-181.

Gibbons, J. D., Olkin, I., and Sobel, M. (1977). Selecting and Ordering Populations: A New Statistical Methodology. New York: John Wiley and Sons, Inc.

Olkin, I., Sobel, M., and Tong, Y. L. (1976). Estimating the true probability of correct selection for location and scale parameter families. Technical Report. No. 110, Department of Statistics, Stanford University, Stanford, California.

Olkin, I., Sobel, M., and Tong, Y. L. (1979). Bounds for a k-fold integral for location and scale parameter models with applications to statistical ranking and selection problems. Technical Report No. 141, Department of Statistics, Stanford University, Stanford, California.

Tong, Y. L. (1978). An adaptive solution to ranking and selection problems. The Annals of Statistics, Vol. 6, pp. 658-672.